DOCUMENT RESUME

ED 084 200 SO 006 619

TITLE	Evaluation Comment, The Journal of Educational Evaluation. Volume 3, Number 4.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
PUB DATE	Dec 72
NOTE	8p.
AVAILABLE FROM	Center for the Study of Evaluation, 145 Moore Hall, University of California, Los Angeles, Los Angeles, California 90024 (Issued Quarterly; Free to professional educators)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Course Evaluation; Curriculum Evaluation; *Educational Objectives; Educational Theories; Evaluation; *Evaluation Criteria; *Evaluation Methods; Formative Evaluation; *Newsletters; Opinions; Summative Evaluation; *Theoretical Criticism
IDENTIFIERS	GFE; *Goal Free Evaluation

ABSTRACT

Each quarterly issue of this journal, available free to professional educators, discusses topics in educational evaluation by presenting articles on evaluation theory, procedures, methodologies, or practices. The topic of the six articles in this newsletter is an examination of free evaluation (GFE). Michael Scriven discusses the role of goal free evaluation in formative and summative evaluation, especially in the evaluation of unintended effects, observes favorable considerations of this method, and presents methodological analogies of GFE in fields other than education. Daniel L. Stufflebeam criticizes Scriven's position and develops four questions he feels to be important in assessing the merit of GFE. Marvin C. Alkin writes that GFE does recognize goals, but that they are wider-context goals rather than specific objectives of a program. W. James Popham proposes that the GFE concept emphasizes results rather than rhetoric and provides a useful caution to educators who are overly enamored of instructional objectives. George P. Kneller argues with the logic of Scriven's argument and finds the issue one of taste rather than of theory. (KSM)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Journal of Educational Evaluation

Evaluation comment

December 1972 Vol. 3, No. 4

Center for the Study of Evaluation
UCLA

ED 084200

SP006 619

PROSE AND CONS ABOUT GOAL-FREE EVALUATION

Michael Scriven

Introduction

In the winter of 1970-71, the National Center for Educational Communications of USOE asked ETS to evaluate the disseminable products of the regional labs and R&D centers. The reward for success was to be substantial grants to assist dissemination. ETS set up an external committee to do the evaluation, under the chairmanship of David Krathwohl, and provided very extensive and excellent staff support for what had to be a rather rapid review. In order to standardize the practice as well as the products of the committee (on which I served) I began to develop a standard form to serve as a check list for us and, when filled out, as a summary for ETS and NCEC. There were originally about 70 entries in what became known as the Product Evaluation Pool, and they ranged from toys for pre-schoolers through publications on teacher training and bilingual curricula, to vast new systems for managing schools. On these, we had varying amounts of data about field trials, mostly very thin, we had the write-ups by the producing staff and other observers, and we had the products themselves. Other input was the list of current USOE priorities in education.

It seemed very natural to start off the evaluation form with a rating of goals of the project and to go on with a rating of the effectiveness in meeting them, costs, etc. By the sixth draft of the form, another item had become very prominent, namely side-effects. Naturally, these had also to be rated, and in one case a product finished up in the Top Ten in spite of zero results with respect to its intended outcomes because it did so well on an unanticipated effect.

—writing on goal-free evaluation

Michael Scriven is a Professor of Philosophy at the University of California, Berkeley

Daniel L. Stufflebeam is the Director of the Evaluation Center and Professor of Education at Ohio State University

Marvin C. Alkin is the Director of the Center for the Study of Evaluation and Associate Professor of Education at UCLA

W. James Popham is the Director of the Instructional Objectives Exchange and Professor of Education at UCLA

George F. Kneller is an educational philosopher and Professor of Education at UCLA

Intended and unintended effects — why distinguish?

Reflecting on this experience later, I became increasingly uneasy about the separation of goals and side-effects. After all, we weren't there to evaluate goals as such—that would be an important part of an evaluation of a *proposal*, but not (I began to think) of a *product*. All that should be concerning us, surely, was determining exactly what effects this product had (or most likely had), and evaluating those, whether or not they were intended.

In fact, it was obvious that the rhetoric of the original proposal which had led to a particular product was frequently put forward as if it somehow constituted supporting evidence for the excellence of the product. This rhetoric was often couched in terms of the "in" phrases of five-year-old educational fads, sometimes given a swift updating with references to the current jargons or lists of educational priorities. That is, the rhetoric of intent was being used as a substitute for evidence of success. Was it affecting us? It would be hard to prove it didn't. And it contributed nothing, since we were not supposed to be rewarding good intentions.

Furthermore, the whole *language* of "side-effect" or "secondary effect" or even "unanticipated effect" (the terms were then used as approximate synonyms) tended to be a put-down of what might well be the crucial achievement, especially in terms of new priorities. Worse, it tended to make one look less hard for such effects in the data and to demand less evidence about them—which is extremely unsatisfactory with respect to the many potentially very harmful side-effects that have turned up over the years.

It seemed to me, in short, that consideration and evaluation of goals was an unnecessary but also a possibly contaminating step. I began to work on an alternative approach—simply, the evaluation of *actual* effects against (typically) a profile of *demonstrated* needs in this region of education. (This is close to what Consumers' Union actually does.) I call this Goal-Free Evaluation (GFE).

Goal-free formative evaluation

At first, it seemed that the proper place for goal-free evaluation (GFE) was in the summative role, like the NCEC activity. In the formative situation, the evaluator's principal task must surely be telling the producer whether the project's goals were being met.

But the matter is not so simple. A crucial function of good formative evaluation is to give the producer a preview of the summative evaluation. Of course, a producer has made the bet that if the goals of the project are achieved, the summative evaluation will be or should be favorable. But one can scarcely guarantee the non-occurrence of undesirable side-effects—and one should not overlook the possibility

FILMED FROM BEST AVAILABLE COPY

of desirable ones that can be cultivated with some care and attention in later developmental cycles. Now, who is going to give the producer a sneak preview of summative results? The staff evaluator will try, and often can do a very good job. But that role is not conducive to objectivity—not only is it dependent on the payroll (and hence one where criticism can produce resentments with which the evaluator will have to live), but it is also very quickly tied in to the production activity. Typically, the staff evaluators are the actual authors of most of the tests in curriculum products, and responsible for some of the form and content of much of the rest. Finally, the staff person is likely to have occupational tunnel-vision with respect to the effects of the materials (or methods, etc.,)—that is, a tendency to look mainly in the direction of the announced goals.

Hence, it now seems to me that a producer or staff evaluator who wants good *formative* evaluation has got to use some external evaluators to get it. Using them does not render the staff evaluator redundant; on the contrary, implementation or correction of the external evaluation depends in large part on the staff person. Psychologically, the staff evaluator may find it priceless to have support from an external source for some personal—and previously unshared—worries or complaints. Now, what I have said so far supports a practice of many producers in using external evaluators. But what I have said also implies—because it springs from the hunt for objectivity/independence—the desirability of arranging goal-free conditions for the external evaluator.

As summative evaluation becomes increasingly goal-free —and I believe it will—the formative evaluation must do so to preserve the simulation. But forget that point; the same conclusion is forced on us by interest in picking up what are for the producer "side-effects." The less the external evaluator hears about the goals of the project, the less tunnel-vision will develop, the more attention will be paid to *looking for actual* effects (rather than *checking* on *alleged* effects).

Other favorable considerations

Look at the effects of considering goals on those who formulate them. It is likely to seem to them that it will pay better to err in the direction of grandiose goals rather than modest ones—as one can see from experience in reading proposals requesting funds, where it's entirely appropriate to evaluate goals. This strategy assumes that a gallant try at Everest will be perceived more favorably than successful mounting of molehills. That may or may not be so, but it's an unnecessary noise source for the evaluator.

The alleged goals are often very different from the real goals. Why should the evaluator get into the messy job of trying to disentangle *that* knot?

The goals are often stated so vaguely as to cover both desirable and undesirable activities, by almost anyone's standards. Why try to find out what was really intended—if anything? (Similarly, the stated goals often conflict—why try to decide which one should supervene.)

A trickier point. The identification of "side-effects" with "unanticipated effects" is a mistake. Goals are only a subset of anticipated effects; they are the ones of special importance, or the ones distinctive of this project. (For example, the goals of a new math curriculum project do not usually include "employing a secretary to type up corrected copy," but of course that effect is anticipated.) Hence, "side-effects" includes more phenomena than "unanticipated effects," and some of the ones it alone includes may be important. In short, evaluation with respect to goals does not even include all the anticipated effects and gives much too limited a profile of the project. Why get into the business of trying to make distinctions like this?

Since almost all projects either fall short of their goals or over-achieve them, why waste time rating the goals; which usually *aren't* what is achieved?

GFE is unaffected by—and hence does not legislate against—the shifting of goals midway in a project. Given the amount of resentment caused by evaluation designs that require rigidity of the treatment throughout, this is an important benefit. But it's a real advantage only to the extent that the project remains within the much larger but still finite ballpark the GFE has carved out of the jungle of possible effects.

Unfavorable considerations — methodological and practical

These are usually an amalgam of criticisms from various sources, sometimes real quotes.

"The GFE'r simply substitutes his own goals for those of the project." No. The GFE may use USOE's goals, or what the best evidence identifies as the needs of the nation, as standards; but simply to use his (or her) own personal preferences would obviously be to invalidate the evaluation. One needs standards of merit for an evaluation, indeed; the error is to think these have to be the goals of the evaluator or the evaluated. Another, commonly connected, error is to think that all standards of merit are arbitrary or subjective. There's nothing subjective about the claim that we need a cure for cancer more than a new brand of soap. The fact that some people have the opposite preference (if true) doesn't even weakly undermine the claim about which of these alternatives the *nation* needs most. So the GFE may use needs and not goals, or the goals of the consumer or the funding agency. Which of these is appropriate depends on the case. But in no case is it proper to use *anyone's* goals as the standard unless they can be *shown* to be the appropriate ones *and* morally defensible.

"Great idea—but hopelessly impractical. You can never keep the evaluator from inferring the goals of the project." This is certainly false; I and others have done evaluations where only the feeblest guesses would be possible, and of no great interest. If you control the data going to the evaluator, you can obviously reduce it to the point where goals are not inferable. And interesting—not exhaustive—evaluations are still possible. An evaluator with considerable experience of goal-based evaluation does indeed find it tempting, in fact almost neurotically necessary, to reach for the security blanket of goals. But once one learns to do without it, then, like riding a bicycle or swimming without the aids one uses at first, there is a remarkable sense of freedom, of liberation.

"Why use an evaluator who only gets part of the data—you simply increase the chance that some of the most important effects (which happen to have been intended) will be missed?" Yes, this is the trade-off. The value of GFE does not lie in picking up what everyone already "knows," but in noticing something that everyone else has overlooked, or in producing a novel overall perspective. Of course, when summative time comes around, the intended effects had better be large enough to be obvious to the unaided (but expert) eye or, in general, they aren't worth very much. (The same is therefore true to a lesser extent for formative evaluation.)

"Attacking the emphasis on careful goal-formulation approaches can only lead to poor planning, a catch-as-catch-can approach, and general carelessness—which you are giving intellectual sanction." Planning and production require goals, and formulating them in testable terms is absolutely necessary for the manager as well as the internal evaluator who keeps the manager informed. That has nothing to do with the question of whether the external evaluator needs or should be given any account of the project's goals.

"I still can't see how GFE is supposed to work in practice. You can't test for all possible effects, and it's surely absurd to think you shouldn't even *bother* with testing the real goals." The external evaluator is not *there* to test goals, but rather to evaluate achievement which turns out to be conceptually distinct—and often different in practice, too. As to the idea that GFE requires testing for every possible effect, the best reply is to say that any evaluator worth hiring has to look for side-effects, and there's no limitation on where or in what form they crop up. So even the goal-based evaluator (GBE'r) has to do this allegedly impossible task. (And so, for that matter, does any applied scientist searching for the effects of a new drug—or the scientist looking for unknown causes of an important effect, e.g., death or cancer: except he searches for every possible cause, not effect.) The GFE'r looks at the treatment and/or curricular materials, after all, and can immediately formulate some hypothesis about probable effects, based on previous experience and knowledge of the research literature. Often, too, the GFE'r can look at the results of quizzes etc., though it's desirable to do that *after* formulating the hypothesis just mentioned, to avoid premature fixation on the variables of concern to the project.

"I'm afraid the GFE is going to be seen as a threat by many producers, perhaps enough to prevent its use." It's true that even GBE was and is so threatening that its introduction has been prevented or rendered useless on many projects. But it has gradually become increasingly a requirement, and the standards for it are creeping upwards. The same is likely to be true of GFE. Now it's important to see why GFE is *more* of a threat. Primarily this is because the GFE'r is less under the control of management; not only are the main variables no longer specified by management, but they may not even *include* those that management has been advertising. The reactions by management to GFE have really brought out the extent to which evaluation has become or has come to seem a controllable item, an unhealthy situation. The idea of an evaluator who won't even *talk* to you for fear of contamination can hardly be expected to make the producer rest easy. It's probably very important, psychologically, to talk to your judge, to feel you've got across a sense of your mission, the difficulties, etc. We all have some faith in "tout comprendre c'est tout pardonner." But the evaluator isn't our judge, just the judge of something we've produced. Even if it's not much good, there's a long way to go before blame can be laid at the producer's door. If a producer really cares about quality control it won't do to insist that the project's definition of quality must be used.

Methodological analogies of GFE (in other fields)

The Intentional Fallacy. In the field of aesthetics it has been widely but not universally accepted that it is fallacious for a critic to consider the intentions of the artist in assessing the work of art. If the "meaning" doesn't show, it doesn't (or shouldn't) count. I am inclined to think this is a perverse view, a purist limit that goes beyond the bounds of sense. The titles of paintings, the locale of photographers, program notes at the symphony, the period of a building, even the biographies of Russian novelists, "cast new light on" the art object itself, and are interesting in themselves. The fallacy is to suppose that the only legitimate framework in which to see a work of art is as an autonomous entity. Art can enlighten, it can give pleasure, it can communicate feeling, and so on—and there's nothing in there that says the background and context of the artwork can't contribute. It's really a case where the consumer can choose. One may say that assessing the *artist* legitimately brings in these considerations, but assessing the *artwork* does not—but the slight attraction of this "tidying-up" move scarcely amounts to a compelling argument for any reasonable man.

In the educational materials production situation, on the other hand, as in the consumer field in general, we can usually establish that the intentions of the producer are of negligible concern to the consumer by comparison with satisfactory performance on the criterion dimensions (e.g., gains in reading scores). Not only *is* this so, but there seems to be little reason why it *shouldn't* be so. When the history of educational R&D is written (if ever historians can be found to stoop to such a low-status task which happens to be socially valuable) then the intentions of producers will be of great interest. For the future producer, a study of these may be far more valuable than a study of the products.

So the "intentional fallacy" is not, in my view, a fallacy in the area where the term was introduced—but it would be one in the evaluation of consumer goods.

Motives and Morality. A tremendous tension has long existed in philosophical ethics between those who believe that the morality of acts is principally determined by their motivation ("He *meant* well") and those who would assess acts in terms of their consequences alone ("Write *that* on his gravestone; first, he should be shot"). Current pop ethics is on the conscience trip—the "pragmatist" is seen as the opposition.

The special feature of this case is that the act involves the motive in a much more intimate way than the product involves the producer's intent. It has been argued that the same physical motions performed with different intentions are definitionally a different act; the distinctions between manslaughter and murder, between borrowing and theft, erring and lying, for example, are said to be distinctions between different acts. One cannot argue that a programmed text supposed to teach economics better than the competition but which actually teaches reading better (and economics the same) is crucially different for the consumer from one in which the side-effect was the primary aim of the producer. And it is for just this reason I prefer the role of the GFE'r for summative evaluation.

On the philosophical issue: I prefer to say that neither exclusive position is defensible, that the issue is resolved one way or the other in *particular* cases where the *point* of the evaluation becomes clear.

Double-Blind Designs. A correspondent writes, "The so-called 'double-blind' medical experiment isn't blind in terms of goal or purpose. A treatment is being tested for its effect on a specific disease. The 'blind' is strictly in terms of the S's or E's knowledge of who is getting what treatment. Thus I think your use of the analogy is inappropriate." The analogy is not intended to be an identity. The point of the analogy is to remind one that medical research, until the scurvy study, ignored the error due to the agent and evaluator knowing that the treatment being given to a particular patient was a dummy. Not only did this affect the agent's behavior in giving it, but it affected the evaluator's care in assessing the effects. After all, how could one seriously look for therapeutic results from a sugar-pill? "Blinding" the assessor made the search equally careful in both cases. Analogously, "blinding" the educational evaluator ensures (to the maximum possible extent?) equal care in looking for effects that happen not to have been goals. Now it's true that the GFE'r may make it the first order of business to infer the goals of the producer. In fact, that's what happened in the second GFE study of which I have received details. (But in the medical case this is often possible, too. In 1958 or so I spent a great deal of time refining placebo effect research designs; the problems of matching for the taste and side-effects of the experimental drug, amongst other difficulties, are typically not solvable.) All one can

do is to make it as hard as possible. In particular, one can try to cut out cues which allow inference of intent other than via noticing success. It's not disastrous if the medical researcher infers *from the results* that treatment B must have been the new medication, treatment A the placebo. The inference may or may not be correct; it can only be damaging if it is made *during* the experiment and hence might influence the later procedures. But even that possibility can usually be handled by splitting the role of recorder from that of agent. By analogy, we cannot get too worried about an evaluator who, seeing massive gain scores on an addition-of-integers test, infers that a major goal of the materials was to improve addition of integers. On the other hand, we must try to avoid having the evaluator come to this conclusion by reading the introduction to the materials, because that is likely to corrupt his later perceptions. When the evaluator devises special instruments for assessing inventory on a parameter that has not previously been tested, we can isolate the role of the agent doing the testing from the role of the scorer, and we can arrange that the scorer does not know the pretests from the posttests, or the experimental group's tests from the control group's tests.

In the early GFE just mentioned, where the evaluator worked diligently to reconstruct the goals, he was doing this by observing various effects which seemed desirable. He concluded that these were probably intended. But the step of inferring goals was totally unnecessary—he could just as well have left the matter by noting the desirable results. Similarly, where he inferred failure (e.g., at teaching the inquiry approach) he could just as well have made no comment, or noted lack of performance in this desirable dimension, from which the evaluend can conclude failure.

Finally, although it is typical of the medical situation that a major parameter is identified in advance, no evaluation of drugs today can avoid the search for side-effects, from the most remote area of the symptom-spectrum. Nor is this obligation restricted to Federal checks; the formative evaluation of drugs requires that the manufacturer run studies that are both double-blind and side-effect sensitive. It would not be difficult to run these evaluations goal-free, but it has little point; given only the characteristics of the patients to be treated, the goal of the treatment would be fairly obvious. In education, the situation is different—more like preventive medicine.

In sum, I think there's an illuminating analogy between the move to double-blind methodology and the (further) move to GFE. The gains from double-blind were not significant in the physical sciences—it was an innovation of great value to medicine. The gains from GFE are not great for medicine—but it is an innovation that may pay off for education.

SHOULD OR CAN EVALUATION BE GOAL-FREE?

Daniel L. Stufflebeam

> Evaluation is... a methodological activity which... consists simply in the gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical ratings, and in the justification of (a) the data-gathering instruments, (b) the weightings, and (c) the selection of goals.
>
> Michael Scriven. The methodology of evaluation. *AERA Monograph Series on Curriculum Evaluation*, Book 1. Chicago: Rand McNally & Company, 1967, pp. 39ff.

In setting forth the above definition of evaluation, Michael Scriven emphasized that evaluators must evaluate goals. The following is a critique of his more recent position that evaluators should pay no attention to goals. In this regard, I will list and respond to four questions that I believe to be important in assessing the merit of goal-free evaluation (GFE).

Question — Should GFE be considered as a possible alternative to existing models of evaluation?

Answer — No. GFE has been proposed as one methodological strategy that can be used to supplement others, including goal-based evaluation (GBE) and the evaluation of goals. This is consistent with Scriven's past practice of analyzing evaluation in order to identify and describe the many kinds of evaluation that evaluators need to be able to perform. In addition to GFE he has proposed formative, summative, intrinsic, payoff, meta, fact-free (with tongue in cheek, I hope), and pathway evaluation. Scriven has not offered any one of these evaluation types, nor all of them collectively, as a theory or model of evaluation. Thus, we should consider GFE in its proper perspective as one strategy that can be used in conjunction with others in evaluation work.

Question — What is the essence of GFE?

Answer — It is to accurately identify effects and determine their importance and quality. That Scriven believes this can best be accomplished by preventing the evaluator from seeing goal statements seems to me both a secondary issue and an empirical question. Perhaps evaluators can be trained not to develop tunnel-vision upon seeing a set of goal statements but to use them as clues for identifying important outcome variables. The main concern is how best to insure that evaluators will identify and properly judge actual results, whether planned or not.

Question — How should GFE be conducted?

Answer — This presently is the rub. Which variables, instruments, extant data, and standards should the evaluator use? When should he gather his data? And how can program people be protected against the potentially arbitrary actions of an inept or unscrupulous goal-free evaluator, especially when he is employed by an external funding agent that may be a bureaucracy with neither a conscience nor a memory?

Presently Dr. Scriven's response seems to be that two goal-free evaluators should operate independently, beginning about midway in a project and continuing to a point after its completion. While this doesn't guarantee good quality and fair evaluation, it at least provides an opportunity to estimate the "error term" involved in GFE.

The problems of gathering data seem far from solution. There are thousands of potentially relevant attainment variables and associated measuring devices, and GFE methodology does not provide much guidance for choosing among them. Goal statements at least provide hypotheses as to what some (NOT ALL) of the variables are. It would seem that system analyses would be helpful, but these also are goal-based.

As to how to judge the GFE results, we encounter a conceptual problem. Scriven suggests that they should be compared with the results of prior needs assessments. This is sound advice, if prior needs assessments were done. But, if needs assessment is the comparison of the real with an ideal, and if the ideal amounts to a prior statement of macro goals, then needs take on their meaning as a function of the discrepancy between an actual situation and prior goal statements. Hence, needs assessments are goal-based and the use of needs asessment data to determine the value meanings of GFE observations is also goal-based. In this respect, the methodological suggestion seems sound, but it raises a question whether GFE can be goal-free. Further, based on Scriven's 1967 definition, evaluation should not be goal-free. The essence of evaluation is value judgments, these are made in relation to standards, and the standards almost always are goals.

Question — Taken in its essential meaning of accurately identifying and properly judging effects, how much can GFE contribute within a broad evaluation framework?

Answer — A great deal.

This type of GFE is the essence of *identifying and judging needs, opportunities, and problems* to serve as a foundation for determining goals. It is also applicable for *identifying and judging alternative program strategies;* solution strategies need to be asessed for their *power* with respect to a wide range of potential desirable impacts—not just those associated with stated goals. Also, through a comprehensive GFE of alternative program strategies one can get a fix on the *tractability* of each of a range of problems and needs, not just the ones associated with the stated goals. GFE is further useful for *identifying and judging a project's effects.* Scriven is absolutely correct that it's unnecessary in identifying outcomes to focus on the stated project objectives. This will be done directly by the goal-based evaluators, and they probably won't have time to search out side-effects.

On the other side of the ledger, GFE will not suffice for meeting accountability requirements. Sponsors pay money so that certain priority needs (goals, if you will) can be met. These needs must be evaluated, and those responsible for meeting them must be judged in terms of their attempts and their achievements and failures. In some cases it is appropriate to penalize one for failing to produce what was needed and what he agreed to produce, especially if the evaluation revealed that the responsible agent did not try to live up to his agreement but instead worked on something more satisfying to him. Such determinations require the use of GBE, although this does not diminish the desirability of GFE.

Within this brief piece I have commented on Michael Scriven's GFE methodological contribution. It fits in with his pattern of analyzing various methodological aspects of evaluation. GFE is not an alternative model of evaluation; rather it is one evaluation strategy. The essence of the strategy should not be to prevent evaluators from seeing goal statements, but to insure that all relevant effects will be accurately identified and properly judged. Conceptually, based on Scriven's own definition and arguments, it is questionable that GFE can or should be goal-free. The strategy is potentially useful, but far from operational and replicable. Because of its promise, I believe that Scriven and others should further develop it, test it, and report back to the profession on the effects of GFE, whatever they turn out to be.

WIDER CONTEXT GOALS AND GOAL-BASED EVALUATORS

Marvin C. Alkin

In this issue of *Evaluation Comment*, Scriven makes some interesting and important points in defense of what he calls the "goal-free evaluator"—the GFE. This term, GFE, is not to be taken literally. The GFE *does* recognize goals (and not just idiosyncratic ones), but they are to be wider-context goals rather than the specific *objectives** of a program. (USOE goals are mentioned as an example.)

In addition to this broader frame of reference the GFE is to be characterized by a scrupulous concern for objectivity. Not only should he refuse to read the program objectives to avoid contamination by the "rhetoric of intent" but he should even decline to talk to the project director.

Insofar as GFEs bring perspective, objectivity, and independence to evaluations they are indeed "a good thing." Manifestly, however, they are not one of the best things in life since they are not *free*. Evaluation costs money; it removes money from program management and implementation funds. Thus, before programs start hiring GFEs we need to discuss what roles are to be played by an internal evaluator (whom Scriven always assumes to exist) and an external evaluator (including the GFE). How does the presence of one affect the activities and responsibilities of the other and to what purpose is each employed?

*The GFE should perhaps be called an OFE (objective-free evaluator) but the unfortunate auditory association with OAF might lead one to think of him as someone who just sits around with no particular purpose in mind.

First, let me point out that in understanding the nature of the evaluation to be performed, the "internal/external" distinction is not nearly as critical as the designation of the decision context to be served by each evaluator. That is, the ultimate purpose of an evaluation is to provide information upon which present or potential decisions are to be made and it is this crucial factor that distinguishes evaluation from research. The nature of the evaluation that will be performed, framed as it is by a particular decision context, will be dependent upon such factors as who hired the evaluator, who receives the evaluation reports, and the nature of the evaluation decision that is to be made (formative, refunding, adoption, etc.). Thus, when Scriven talks about an "internal evaluator" I presume that he is referring to an evaluator hired by the project director primarily to provide formative information for program modification purposes and whose reports will be directed primarily toward the project director (and perhaps secondarily to the sponsoring agency). In addition to this internal evaluator there should perhaps be an evaluator *hired by* the Superintendent of Schools to report on the project; perhaps the sponsoring agency should also hire an external GBE to report to them. There are many decision contexts requiring evaluation information and it is necessary to establish priorities on these various evaluation requirements.

By "goal-free" Scriven simply means that the evaluator is free to choose a wider context of goals. By his description

he implies that a goal-free evaluation is *always* free of the goals of the specific program and *sometimes* free of the goals of the program sponsor. In reality, then, goal-free evaluation is not really goal-free at all but is simply directed at a different and usually wider decision audience. The typical goal-free evaluator must surely think (especially if he rejects the goals of the sponsoring agency) that his evaluation will extend at least to the level of "national policy formulators." The question is whether this decision audience is of the highest priority in our present concerns for improving evaluation.

The high priority that Scriven attaches to the goal-free evaluator seems to be based primarily upon his experience in considering the evaluation of packaged instructional products designed to be used widely and in a variety of contexts. Scriven's major examples come from product evaluations performed at only a limited number of centers, laboratories, and other organizations that produce validated instructional materials. Each of these organizations has an internal evaluation staff. But the materials they are concerned with represent merely the tip of a giant iceberg of instructional products most of which undergo little or no evaluation—neither formative nor summative nor goal-based of any kind. Moreover, when one considers problems related to the evaluation of instructional *programs* (such as the Title programs—I, III, VII, VIII) and the evaluation of *teachers* (such as that mandated by the Stull Bill in California), then the iceberg of instructional product evaluation pales in importance compared to the Arctic sea of evaluation problems. Thus, while it is difficult to dispute Scriven's point that there is a role for a person called a goal-free evaluator, one can certainly question his judgment as to the areas of greatest "demonstrated need" in evaluation at this time. And if one can question a goal-free evaluator on how well he interprets "demonstrated need," what else is left?

And so, what are the alternatives to a goal-free evaluator? Scriven comes to see the need for goal-free evaluators because he questions the goals (or objectives) specified by project personnel as potentially not being expressions of "demonstrated need" or as being ambiguously stated. If this is the case, then why must the evaluator wait for the program to become fully implemented before providing evaluative feedback on the rightness of goals. In part, this lack of foresight attributed to a goal-based evaluator (GBE) by Scriven is related to his rather limited definition of the role of the evaluator. Scriven thinks of the evaluator as participating in formative and summative evaluation, in essence limiting the evaluative engagement to the period following the adoption of the educational program. This oversight is corrected in the evaluation model of the Center for the Study of Evaluation in which we conceive of the evaluative responsibility beginning with "needs assessment". In the needs assessment stage the evaluator assists in providing explicit data as to the relevance of stated goals to real and demonstrated needs. Scriven's goal-free evaluation is in essence a retrospective (and non-explicit) needs assessment. This would be all right, but for the fact that performing this function retrospectively raises the cost enormously, not only of the evaluation but of a program that may have gone astray and which could have been brought back on course at an earlier time.

If the goals that are alleged are not the "real" ones or the "right" ones then let the GBE establish a procedure, an explicit procedure, for determining the goals. If mere rhetoric constituted the supporting evidence then let the GBE do a better job in *assessing* the goals. Condemning the GBE procedure because of inadequacies in its execution does not solve the problem. Performing a better job of GBE does offer some hope.

RESULTS RATHER THAN RHETORIC

W. James Popham

Whether Michael Scriven ever uttered the phrase "results rather than rhetoric" I am not certain. I came away from a conference many months ago in Colorado thinking that he had. It was there that Michael was testing an early conception of his goal-free evaluation position. If he didn't use that particular phrase, he probably won't be too displeased if I attribute it to him. After all, not only is the phrase alluringly alliterative, but it conveys a commitment to empirical evidence and a dismissal of mere word wizardry. And Michael Scriven has a strong allegiance to empirical methods and a special flair for walloping word wizards. I don't think he'd mind the attribution at all.

But beyond questions of its ancestry, the idea of results rather than rhetoric, as embodied in Scriven's goal-free evaluation writings, provides a useful caution to those educators who have recently become so enamored of instructional objectives that they think the mere act of articulating their goals precisely is not only the beginning but the end of the instructional ball game. And as you can learn from any baseball pitcher who has set out in the first inning to pitch a shutout, the game's final score is the thing that counts, not good intentions. Goal-based evaluation has offered educators a way of counteracting the heavy emphasis on instructional process which has been so fashionable in our country for years. GBE made it easier to describe intended instructional effects, then see if they were actually produced. But, as Professor Scriven's goal-free evaluation paper reminds us, GBE has often led to a tunneling of vision so that important results of instruction were overlooked. If GFE does nothing more than remind educators to appraise an educational undertaking on the basis of all its important effects, not just those which were described beforehand (even in flawlessly fashioned behavioral objectives), then GFE will have been a useful contribution.

But while the logic of Scriven's GFE stance is commendable, there are a few implementation operations which currently vex me. It's so early in the GFE game that Professor Scriven hasn't had time to wrestle with all of them. He undoubtedly will in time.

First, there was a clear implication in several of his early essays on GFE that the GFE'r could derive special raptures from spotting the educational catastrophes that a goal-blinded evaluator would not discern. Scriven spent a fair amount of time describing how the GFE'r would "set snares" to pick up a program's effects. While discovering *all* important effects are the proper province of the GFE'r, one had the distinct impression that his real kicks came from isolating an undiagnosed malignancy. We'll have to see whether goal-free evaluators can be trained so that they develop a balanced search for the beneficial as well as the harmful results of an instructional program.

Second, there is a practical difficulty which the GFE'r will have trouble resolving, particularly in a formative context. If it is true, as Scriven contends, that actual effects must be evaluated against "a profile of *demonstrated* needs," then clearly the GFE'r will either have to conduct some sort of an independent needs assessment or must rely on an existing effort to demonstrate needs. Relying on an existing needs assessment operation, particularly if carried out by the staff of the project being evaluated, carries with it the same deficits as GBE; that is, there may be subtle project staff biases operating which distort the validity of the assessment. But conducting an independent needs assessment is costly business and may not be considered cost-effective by the project's management. These problems may be more easily resolved in the summative context, because the stakes are often perceived as higher and a summative evaluator may therefore more readily be able to demand the resources needed to secure an unbiased needs profile. But for a formative evaluator, I think this is a sticky problem. We want to foster as much independence for our GFE'r as possible, yet a totally independent needs assessment seems uneconomical.

A third problem stems from the degree to which a GFE'r can remain insulated from the instructional designer's goal preferences when it comes to devising the measures required to assess program impact on learners. In the abstract it is easy for a GFE'r to turn off the instructional designer who is about to spout goal talk. In constructing tests, observation scales, unobtrusive measures, etc., the GFE'r needs to have some kind of clues regarding what results the instruction is *apt* to yield. But as the requisite inferences are made from instructional procedures, materials, etc., there will be a strong likelihood that the project goals will insinuate themselves in the perceptions of the GFE'r. I suspect, therefore, that the possibility of keeping GFE *completely* uncontaminated by goal preferences is unrealistic. We must make it as goal-free as we can.

A final problem with GFE is that many educators—terrorized by the possible repercussions of goal-based evaluation—will use GFE as a philosopher-approved excuse for chucking out goals altogether. Yet Scriven makes it very clear that goals are required for planning, production, and internal evaluation. We must guard against those who will try to use GFE as an intellectually respectable cover for not thinking rigorously about their educational intentions.

Goal-free evaluation is destined to become very popular among educational folk. It is new. It was sired by an eminent academic philosopher who, all blessings abound, speaks with an educated British accent. I can see future evaluators clamoring for specially designed GFE blinders to protect them from the taint of project goals. Short courses in snare-setting will be conducted jointly by university departments of education and state game commissions. GFE will be IN.

But, because I have been persuaded by an eminent academic philosopher who speaks with an educated British accent, I'll have to wait until all this GFE stuff has been tried out in a good number of real educational evaluations. You see, I've recently become somewhat committed to results rather than rhetoric.

GOAL-FULL EVALUATION

George F. Kneller

Professor Scriven advocates goal-free evaluation as a remedy for certain weaknesses in contemporary research design. The remedy, however, is unnecessary, since, as I shall point out, these weaknesses can be corrected more efficiently by modifying either the design itself, or the training of evaluators, or both.

Scriven's most substantial argument in favor of goal-free evaluation is that the more an evaluator concerns himself with the goals of a project, the less likely he is to notice the project's side-effects. This tendency, however, may be corrected in two ways without resort to GFE: first, by training evaluators to observe *both* goals (and outcomes) *and* side-effects; second, by the researcher's specifying as many likely side-effects as possible within the original research design. Thus the researcher himself gathers many of the relevant data while conducting his own project.

Scriven also argues that the use of GFE makes it harder for an evaluator to persuade himself that the goals of a project have been achieved simply because they have been set. But bias in favor of goals is only one of many biases to which evaluators are subject, and little is gained by seeking to eliminate this form of subjectivity while leaving other forms untouched. The wisest course is not to rely on GFE to eliminate one form of subjectivity but to train evaluators in advance to be objective judges in as many respects as possible.

Scriven's other arguments in favor of GFE carry little weight:

- He maintains that research projects often are designed to attain grandiose goals which distract the evaluator's attention from the project's actual achievements. I reply that (a) evaluators should be trained to spot and to criticize grandiose goals, and (b) researchers should be trained to set realistic goals.

- Similarly Scriven asserts that the alleged goals of a project often differ from the real ones. Once again, however, proper training should (a) correct this sort of misunderstanding in the researcher, and (b) improve the evaluator's ability to recognize the discrepancy when it occurs.

- Scriven also claims that the goals of many research designs are too vague. Indeed, they may be. But the way to eliminate the fault is not to introduce GFE after the event but to educate researchers to draw up their designs more carefully at the outset.

- Scriven calls for GFE on the grounds that projects often fail to achieve their goals. But unless we take these goals into account, we shall never know which projects have succeeded in their aims and which have not.

The frailties which Scriven correctly criticizes in research designers may also be found in evaluators, goal-free or otherwise. These are human frailties, and they may come into play anywhere in the course of a project from its preliminary drafting to its completion. It is not enough, therefore, to provide one particular safeguard by introducing GFE after the project is finished. Instead, safeguards should be built into the design at many points.

Also, Scriven makes no provision for the defense of researchers against the bias of his breed of evaluators. In my view, researchers are fully entitled to object to evaluators

who are not concerned to find out what the goals are of the projects they are examining. I am not arguing for a balance of power between researchers and evaluators, but I am saying that evaluators, too, can become irrational and immoderate.

Scriven's metaphors and analogies are bright and amusing, but they do not make for tight logical argument. Moreover, some of them are ill-conceived. The goal-free evaluator, he says, once having learned to be free of a security blanket of goals, "like riding a bicycle or swimming without the aids one uses at first, (experiences) a remarkable sense of freedom, of liberation." In fact, (a) aids are not in the same class as goals, and (b) security blankets may save the lives of those who feel so free that they outswim (lose control of) themselves.

If I had to make a choice, I would reject GFE, not only in consequence of the counter-arguments I draw above but also because of my gestaltist tendency to see things in more or less complete patterns, hence to take the goals of enterprises into consideration. In any case, I do not see the issue as theoretical but as one appealing to taste, about which there can be no dispute. One simply makes a choice and, if called upon to justify the choice, can offer only personal opinions.

Scriven's essay is not without nuggets of wisdom. For example, he points out: "But in no case is it proper to use *anyone's* goals as the standard unless they can be *shown* to be the appropriate ones *and* morally defensible." This remark should be taken seriously not only by researchers and evaluators but by people in every walk of life.

EVALUATION COMMENT, THE JOURNAL OF EDUCATIONAL EVALUATION, **is published by the Center for the Study of Evaluation, University of California, Los Angeles. CSE, which is one of eight educational research and development centers in the country, was founded in June, 1966, by the U.S. Office of Education under the Cooperative Research Act. It is the only federally funded research and development center working exclusively on problems in educational evaluation.**

Each issue of the journal will discuss topics in educational evaluation by presenting articles on evaluation theory, procedures, methodologies, or practices. For each issue of the journal, CSE will request contributions on a specific topic in evaluation from recognized experts in the field; unsolicited manuscripts will not be accepted.

Evaluation Comment **has a current circulation of over 8,000 readers. Each scholar, researcher, or practitioner on our mailing list receives a free copy of the journal. Where additional copies are needed readers are encouraged to reproduce the *Comment* themselves. To be placed on our mailing list, please write to:**

James Burry, Managing Editor
Evaluation Comment
Center for the Study of Evaluation
145 Moore Hall
University of California, Los Angeles
Los Angeles, California 90024

University of California

The Journal of Educational Evaluation
Evaluation comment
Center for the Study of Evaluation
UCLA

Center for the Study of Evaluation
145 Moore Hall
405 Hilgard Avenue
Los Angeles, California 90024

NON-PROFIT ORG.
U.S. POSTAGE
PAID
LOS ANGELES, CALIF.
PERMIT NO. 12378